MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2015/2016

TNL3221 – NATURAL LANGUAGE PROCESSING

(All Sections / Groups)

11 MARCH 2016
9.00 a.m. – 11.00 a.m.
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This question paper consists of 4 pages with 6 questions only.
2. Attempt **FIVE** out of SIX questions only. All questions carry equal marks and the distribution of the marks for each question is given.
3. Please write all your answers in the Answer Booklet provided.

Question 1

(a) Briefly explain the FIVE levels of language processing. [5 marks]

(b) Write a regular expression for each of the following:

i. That matches any word begins with '*p*', followed by at most two vowels (a, e, i, o, u), and ends with '*t*'. [1 mark]

ii. That matches any word begins with '*s*' and ends with '*g*'. [1 mark]

iii. That matches the words: *a*, *an*, and *the*. [1 mark]

iv. That matches an arithmetic expression using integers, addition, and multiplication, such as 2*3+8, 5+9*1. [2 marks]

Question 2

(a) Given the regular expression /hel*lo/:

i. Draw a finite state automation. [3 marks]

ii. Draw a state-transition table for the automaton. [3 marks]

iii. Justify whether it is a deterministic finite state automaton (DFSA) or non-deterministic finite state automaton (NFSA). [1 mark]

(b) When we use a non-deterministic finite state automaton to recognize a string, we might take the wrong path when there is more than one choice at some points. Describe the THREE solutions to the above state problem of non-determinism. [3 marks]

Question 3

(a) Given the following context-free grammar:

$$S \rightarrow NP\ VP$$

$$NP \rightarrow Pronoun \mid Proper\text{-}Noun \mid Det\ Nominal$$

$$Nominal \rightarrow Nominal\ Noun \mid Noun$$

$$VP \rightarrow Verb \mid Verb\ NP \mid Verb\ NP\ PP \mid Verb\ PP$$

$$PP \rightarrow Preposition\ NP$$

Use the shift-reduce approach to search this string: '*They watch a movie in the cinema*'. [7 marks]

(b) How many token, type, bigram, and trigram (including punctuation) are there in the following sentence? [3 marks]

> By providing more natural human machine interfaces, and more sophisticated access to stored information, language processing has come to play a central role in the multilingual information society.

Continued...

Question 4

(a) Given the following unigram and bigram counts:

I	Want	To	Eat	Chinese	Food	Lunch	Thai
3437	1215	3256	938	213	1506	459	315

Table 1: Unigram Counts

	I	Want	To	Eat	Chinese	Food	Lunch	Thai
I	8	1087	0	13	0	0	0	0
Want	3	0	786	0	6	8	6	8
To	3	0	10	860	3	0	12	5
Eat	0	0	2	0	19	2	52	28
Chinese	2	0	0	0	0	120	1	0
Food	19	0	17	0	0	0	0	0
Lunch	4	0	0	0	0	1	0	0
Thai	3	0	0	0	0	150	2	0

Table 2: Bigram Counts

i. Compute the bigram probabilities P(Want|I), P(To|Want), P(Eat|To), P(To|Eat), P(Lunch|Eat), and P(Eat|Lunch). [Round your answers to two decimal places.] [3 marks]

ii. Suppose P(I|<s>) = 0.25 and P(</s>|Lunch) = 0.35, calculate the probability of the sentence "I Want To Eat Lunch". [Round your final answer to five decimal places.] [5 marks]

iii. Compute the bigram probabilities P(Chinese|Eat) and P(Thai|Eat). [Round your answers to two decimal places.] Briefly explain your interpretation of the probabilities obtained. [2 marks]

Question 5

(a) Draw the architecture of a dialogue system. [5 marks]

(b) 'Grounding' property of human conversation is applied in the stated conversation between both speaker A and hearer C.

C_1: … I need to travel in May.
A_1: And, what day in May did you want to travel?
C_2: OK uh I need to be there for a meeting that's from the 12^{th} to the 15^{th}.
A_2: And you're flying into what city?
C_3: Seattle.
A_3: And what time would you like to leave Pittsburgh?
C_4: Uh hmm I don't think there's many options for non-stop.
A_4: Right. There's three non-stops today.

Continued…

i. Describe the purpose of 'grounding' in human conversation. [2 marks]

ii. What is the usage of using '*And*' at the beginning of each conversation by A_1, A_2 and A_3? [1 mark]

iii. What is the 'acknowledgement' token has used in the conversation by C_2 and C_4? [2 marks]

Question 6

(a) Discuss the relative advantages and disadvantages of top-down and bottom-up parsing. [4 marks]

(b) Use Penn Treebank tagset to tag each word of this sentence: [6 marks]

A tagger processes words and attaches a tag to each word.

1.	CC	Coordinating conjunction	25.	TO	*to*
2.	CD	Cardinal number	26.	UH	Interjection
3.	DT	Determiner	27.	VB	Verb, base form
4.	EX	Existential *there*	28.	VBD	Verb, past tense
5.	FW	Foreign word	29.	VBG	V, gerund/pres. particle
6.	IN	Preposition/sub. conjunction	30.	VBN	V, past particle
7.	JJ	Adjective	31.	VBP	V, non-3rd ps. sing. pres.
8.	JJR	Adjective, comparative	32.	VBZ	V, 3rd-ps. sing. present
9.	JJS	Adjective, superlative	33.	WDT	*wh*-determiner
10.	LS	List item marker	34.	WP	*wh*-pronoun
11.	MD	Modal	35.	WP$	Possessive *wh*-pronoun
12.	NN	Noun, singular or mass	36.	WRB	*wh*-adverb
13.	NNS	Noun, plural	37.	#	Pound sign
14.	NNP	Proper noun, singular	38.	$	Dollar sign
15.	NNPS	Proper noun, plural	39.	.	Sentence final punct.
16.	PDT	Predeterminer	40.	,	Comma
17.	POS	Possessive ending	41.	:	Colon, semicolon
18.	PRP	Personal pronoun	42.	(	Left bracket character
19.	PPS	Possessive pronoun	43.	)	Right bracket character
20.	RB	Adverb	44.	"	Straight double quote
21.	RBR	Adverb, comparative	45.	‘	Left open single quote
22.	RBS	Adverb, superlative	46.	“	Left open double quote
23.	RP	Particle	47.	’	Right open single quote
24.	SYM	Symbol	48.	”	Right open double quote

Table 3: Penn TreeBank Tagset

End of paper